

Approximate Inference and Stochastic Optimal Control

Konrad Rawlik¹, Marc Toussaint², and Sethu Vijayakumar¹

¹Statistical Machine Learning and Motor Control Group, University of Edinburgh

²Machine Learning and Robotics Group, TU Berlin

September 22, 2010

Abstract

We propose a novel reformulation of the stochastic optimal control problem as an approximate inference problem, demonstrating, that such a interpretation leads to new practical methods for the original problem. In particular we characterise a novel class of iterative solutions to the stochastic optimal control problem based on a natural relaxation of the exact dual formulation. These theoretical insights are applied to the Reinforcement Learning problem where they lead to new model free, off policy methods for discrete and continuous problems.

Contents

1	Introduction	2
2	Preliminaries	2
3	On Stochastic Optimal Control and KL divergences	4
3.1	Bayesian Model of Control Problems	4
3.2	General Duality	5
3.3	Iterative Solution	6
3.3.1	Finite Horizon Case	7
3.3.2	Infinite Horizon Case	8
3.4	Relation to Previous Work	9
3.4.1	Approximate Inference Control	9
3.4.2	Path Integral and KL control	9
3.4.3	Expectation Maximization Approaches	11
3.5	Conclusion	12
4	Reinforcement Learning	12
4.1	Finite Problems	12
4.1.1	Relation to Classical Algorithms	13
4.1.2	Results	13
4.2	Continuous problems	14
4.2.1	The LS Ψ algorithm	14
4.2.2	Results	15
4.3	Conclusion	17
A	Supplementary proofs	20

1 Introduction

In recent years the framework of *stochastic optimal control* (SOC) [27] has found increasing applicability in the domain of planning and control of realistic robotic systems [17, 35] while also finding widespread use as one of the most successful normative models of human motion control [32, 9]. In general SOC can be summarised as the problem of controlling a stochastic system so as to minimise expected cost. The general problem subsumes a variety of different problems all based on slightly different assumptions, e.g. Markov Decision Processes [29], Reinforcement Learning [28] or Adaptive Control. The increased use of the general formalism in high dimensional and non linear settings necessitates the development of novel efficient methods, while it's diverse nature makes novel theoretical insights into the general problem extremely desirable.

In the most general setting, the stochastic optimal control problem with arbitrary dynamics and cost function is analytically intractable and significant previous research has focused on developing efficient *approximate* solution methods [10, 15]. In particular there have been, in recent years, an increasing number of attempts to relate the stochastic optimal control problem to problems from the domain of probabilistic inference, specifically maximum likelihood problems, e.g., [34, 2], and inference problems [12, 33]. The hope was that by finding such correspondences, the large number of available efficient Machine Learning [4] approaches will become applicable to the stochastic optimal control problem.

In this paper we propose a reformulation of the general stochastic optimal control problem as a problem of approximate probabilistic inference. Unlike previous theoretical work on this issue [13, 31, 12, 18] this reformulation is exact without making further assumptions, though this comes at the cost of a lack of a closed form solution. However, the exact reformulation of stochastic optimal control as an inference problem is, in itself, not the main motivation of this work. Rather we see it as a starting point for development of novel approaches to the problem, which draw from the alternative interpretation. We show for example that the reformulation can be directly related to the previously proposed approximate inference control framework [33] which allows us to clarify the relation of the latter to stochastic optimal control.

Importantly we demonstrate that a, in the context of a probabilistic interpretation, natural relaxation of the new formulation directly leads to a novel class of iterative solutions for the stochastic optimal control problem. We characterise the form of these iterations and highlight their relation to previous applications of Expectation Maximisation algorithm in this area [34, 2]. We also directly demonstrate the applicability of these results, by deriving novel model free, off policy Reinforcement Learning algorithms for discrete and continuous problems.

We would like to note that this text forms part of the first author's PhD progress report (May 2010) submitted to the University of Edinburgh Graduate School. This document aims to make this work available to a wider audience as we have been made aware of the recent work by Kappen et.al. [1] (pursued independently and in parallel) which shows distinct parallels to the methods developed here, with specifically the *Dynamic Policy Programming* (DPP) algorithm having significant overlap with the here proposed LS Ψ algorithm, although the motivation and derivation differ. Furthermore we claim that the results presented here go beyond the work of [1] by providing a more general framework, relating it to previous approaches in Stochastic Optimal Control and Reinforcement Learning, and by demonstrating applicability of the algorithm to continuous problems. In particular we highlight a class of approximations which lead to analytical expressions in the continuous setting, mitigating the need to use computationally expensive numerical or Monte Carlo methods anticipated by [1].

The remainder of this paper is structured as follows. After introducing necessary concepts of stochastic optimal control in section 2 we present in section 3 our theoretical results relating to the approximate inference formulation of stochastic optimal control problems. These are then applied in section 4 to the Reinforcement learning problem.

2 Preliminaries

In the remainder of this text we will consider control problems which can be modeled by a *Markov Decision Process* (MDP) and before proceeding we first recall the standard formalism. We shall keep this exposition

rather brief, only introducing concepts necessary for the development of the theory and methods in this paper. For a broader review one may refer to the 1st Year proposal or [29], or for a more thorough treatment to any of the numerous text books on the subject, e.g., [27, 28, 3].

A MDP provides in general a model of a sequential decision process, where an agent observes it's state, chooses a control and then transitions to a new state whilst incurring a certain cost. More formally, let $x_t \in \mathcal{X}$ be the state and $u_t \in \mathcal{U}$ the control signals at times $t = 1, 2, \dots, T$. In order to simplify notation we will denote whole state and control trajectories $x_{1..T}, u_{0..T}$ by \bar{x}, \bar{u} . Let $P(x_{t+1}|x_t, u_t)$ be the transition probability for moving from x_t to x_{t+1} under control u_t and let $\mathcal{C}_t(x, u) \geq 0$ be the cost incurred for choosing control u in state x at time t . A policy for time step t , $\pi_t(u_t|x_t)$, is the conditional probability of choosing the control u_t given the state x_t . In the interest of a less cluttered notation we shall in the following in general drop the subscript t on π if it is obvious from the context. An important family of policies is the set \mathcal{D} of *deterministic policies*, which are policies given by a conditional delta distribution, i.e. $\pi(u_t|x_t) = \delta_{u_t=\tau(x_t)}$ for some function τ . The stochastic optimal control problem consists of finding a deterministic policy¹ which minimises the expected cost, i.e., solving

$$\pi^* = \underset{\pi \in \mathcal{D}}{\operatorname{argmin}} \left\langle \sum_{t=0}^T \mathcal{C}_t(x_t, u_t) \right\rangle_{q_\pi}, \quad (1)$$

where

$$q_\pi(\bar{x}, \bar{u}|x_0) = \pi(u_0|x_0) \prod_{t=1}^T \pi(u_t|x_t) P(x_{t+1}|x_t, u_t), \quad (2)$$

is the distribution over trajectories with start state x_0 and under policy π .

In the case of an infinite time horizon, i.e. for $T \rightarrow \infty$, we will restrict ourselves to the discounted cost formulation. That is we will assume the cost to be a discounted time stationary cost, so that $\mathcal{C}_t(x_t, u_t) = \gamma^t \mathcal{C}(x_t, u_t)$ for some discount factor $\gamma \in [0, 1]$.

For a given policy π we may define the value function $\mathcal{J}_t^\pi : \mathcal{X} \rightarrow \mathbb{R}$, as the mapping from a state x to the expected cost of starting in x at time t and following π thereafter, i.e.,

$$\mathcal{J}_t^\pi(x) = \left\langle \sum_{k=t}^T \mathcal{C}_k(x_k, u_k) \right\rangle_{q_\pi(x_{t+1} \dots T, u_{t+1} \dots T | x_t = x)}. \quad (3)$$

Similarly we may, for a given policy π , define the state-control, or state-action as it is more commonly known, value function $\mathcal{Q}_t^\pi : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$, which for a given x, u gives the expected cost of starting in state x at time t , choosing control u and following π thereafter, i.e.,

$$\mathcal{Q}_t^\pi(x, u) = \left\langle \sum_{k=t}^T \mathcal{C}_k(x_k, u_k) \right\rangle_{q_\pi(x_{t+1} \dots T, u_{t+1} \dots T | x_t = x, u_t = u)}. \quad (4)$$

Of obvious interest are the value and state action value functions of π^* , which we denote by \mathcal{J}^* , \mathcal{Q}^* . They are sufficient, in the case of \mathcal{J}^* together with knowledge of the transition probability and cost function, to characterise the optimal policy. An equation of particular importance in this context is the Bellman optimality equation

$$\mathcal{J}_t^*(x_t) = \min_{u_t \in \mathcal{U}} \left[\mathcal{C}_t(x_t, u_t) + \int_{x_{t+1}} P(x_{t+1} | x_t, u_t) \mathcal{J}_{t+1}^*(x_{t+1}) \right], \quad (5)$$

which in the infinite horizon discounted cost setting gives the following fixed point equation for the optimal value function,

$$\mathcal{J}^*(x) = \min_{u \in \mathcal{U}} \left[\mathcal{C}(x, u) + \gamma \int_y P(y | x, u) \mathcal{J}^*(y) \right]. \quad (6)$$

¹n.b. it can be shown that for problems of the type described here there exists a optimal policy which is deterministic [29]

Although the Bellman equations are in general not analytically tractable, in either the finite or infinite horizon case, they have provided the starting point for a large number of approaches for solving the stochastic optimal control problem, and will indeed be closely related to the starting point of the formulation proposed in this paper.

As an aside we note that throughout this paper we will in general be working under the more general assumption of infinite control and state spaces and hence use integrals, as has been already done in the Bellman equations. This is done with the understanding that for discrete problems these simply reduce to finite sums.

3 On Stochastic Optimal Control and KL divergences

We will now state our main theoretical results which will form the basis of the work presented in section 4 and proposed future work. Specifically we will show how stochastic optimal control can be formulated as a approximate inference problem in a certain probabilistic model. For the purpose of this paper, we define *approximate inference*, as the approximation of a true posterior within some family of distributions by minimization of some divergence measure. The divergence measure which we will consider here is the Kullback-Leibler divergence, which, for two distributions q & p over \mathcal{X} , is defined as

$$\text{KL}(q\|p) = \int_{\mathcal{X}} q(x) \log \frac{q(x)}{p(x)} . \quad (7)$$

After introducing the probabilistic model in subsection 3.1, we will, in subsection 3.2, state and discuss our general duality result. As this result does not directly lead to a closed form solution of the stochastic optimal control problem we will then proceed to demonstrate in subsection 3.3 that under a relaxation of the exact dual a novel class of iterative approaches arises, which allows for closed form iterations. We will then derive such iterations for the finite and infinite horizon case. Finally we will discuss the relations of these results to previous work in the field.

3.1 Bayesian Model of Control Problems

In most general terms, we would define inference based control in terms of a Dynamic Bayesian Network which includes multiple state, task, and control variables in each time slice. We would distinguish three types of random variables, state and control variables, defined as in the stochastic optimal control framework, and additionally a set of variables, which we will refer to as *task variables*, which capture the achievement of the objective described by the cost. In general the states and controls are latent variables and we wish to marginalise the states and infer the controls. The task variables on the other hand are observed, in the sense that we aim to make inference about the controls in the case of an achieved task.

More formally the model takes the form illustrated by the graphical model in figure Figure 1. We relate the task likelihood to the classical cost by choosing

$$P(r_t = 1 | x_t, u_t) = \exp\{-\mathcal{C}_t(x_t, u_t)\} , \quad (8)$$

which is well defined due to the restriction $\mathcal{C}_t(\cdot, \cdot) \geq 0$. The complete joint is now given by

$$P(\bar{x}, \bar{u}, \bar{r} | x_0; \pi) = q_{\pi}(\bar{x}, \bar{u}) \prod_{t=0}^T P(r_t | u_t, x_t) , \quad (9)$$

where q_{π} , the trajectory distribution under a policy, has been defined previously in (2). As indicated, our main interest will be with the posterior under the *assumed* observation of task success, and we will use the notation

$$p_{\pi}(\bar{x}, \bar{u} | x_0) = P(\bar{x}, \bar{u} | x_0, \bar{r} = 1; \pi) = \frac{1}{P(\bar{r} = 1 | x_0; \pi)} q_{\pi}(\bar{x}, \bar{u}) \prod_{t=0}^T P(r_t = 1 | u_t, x_t) . \quad (10)$$

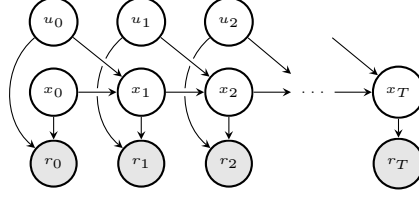


Figure 1: The graphical model of for the Bayesian formulation of the control problem in the finite horizon case. In the infinite horizon case we obtain a stochastic markov process.

We would like to note that we have presented a model of sufficient structure for the results which follow, with the understanding that in cases with additional structure, better algorithm and stronger results may be obtainable. In particular we make no further assumption about conditional independence structure which is often present between subsets of state and control variables. Furthermore we view a single task variable in each time step as a sufficient representative of any set of task variables one might conceive, e.g., endeffector targets variables, collisions etc.. Eventually, the observed task variables only induce extra potentials on the remaining state and control variables. Therefore one could avoid introducing them at all in the formalism. However, we find their notion helpful to develop the theory.

3.2 General Duality

We may now directly state the main result relating the discussed Bayesian model to stochastic optimal control.

Proposition 1 (Rawlik & Toussaint [22]). *Let π^0 be an arbitrary stochastic policy and \mathcal{D} the set of deterministic policies, then the problem*

$$\operatorname{argmin}_{\pi \in \mathcal{D}} \text{KL}(q_\pi \| p_{\pi^0}) \quad (11)$$

is equivalent to the stochastic optimal control problem with cost per stage

$$\hat{\mathcal{C}}_t(x_t, u_t) = \mathcal{C}_t(x_t, u_t) - \log \pi^0(u_t | x_t) \quad (12)$$

Proof. Let $\pi_t(u_t | x_t) = \delta_{u_t = \tau_t(x_t)}$, for some function τ , then

$$\begin{aligned} \text{KL}(q_\pi \| p_{\pi^0}) &= \log P(\bar{r} = 1) + \int_{\bar{x}} \int_{\bar{u}} q_\pi(\bar{x}, \bar{u}) \log \frac{q_\pi(\bar{x}, \bar{u})}{q_{\pi^0}(\bar{x}, \bar{u})} \\ &\quad + \int_{\bar{x}} \int_{\bar{u}} q_\pi(\bar{x}) \pi(\bar{u} | \bar{x}) \sum_{t=0}^T \log \frac{1}{\exp\{-\mathcal{C}_t(x_t, u_t)\}} \end{aligned} \quad (13)$$

$$= \log P(\bar{r} = 1 | x_0; \pi^0) + \text{KL}(q_\pi(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) + \int_{\bar{x}} \int_{\bar{u}} q_\pi(\bar{x}) \delta_{\bar{u} = \tau(\bar{x})} \sum_{t=0}^T \mathcal{C}_t(x_t, u_t) \quad (14)$$

$$= \log P(\bar{r} = 1 | x_0; \pi^0) + \text{KL}(q_\pi(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) + \int_{\bar{x}} q_\pi(\bar{x}) \sum_{t=0}^T \mathcal{C}_t(x_t, \tau_t(x_t)) . \quad (15)$$

Furthermore the divergence between the controlled process, q_π , and prior process, q_{π^0} is

$$\text{KL}(q_\pi(\bar{x}, \bar{u}) \| p_{\pi^0}(\bar{x}, \bar{u})) = \int_{\bar{x}} \int_{\bar{u}} q_\pi(\bar{x}, \bar{u}) \sum_{t=0}^T \log \frac{\delta_{u_t = \tau_t(x_t)}}{\pi^0(u_t | x_t)} \quad (16)$$

$$= - \int_{\bar{x}} q_\pi(\bar{x}) \sum_{t=0}^T \log \pi^0(\tau_t(x_t) | x_t) . \quad (17)$$

Hence,

$$\text{KL}(q_\pi \| p_{\pi^0}) = \log P(\bar{r} = 1 | x_0; \pi^0) + \left\langle \sum_{t=0}^T [C_t(x_t, \tau_t(x_t)) - \pi^0(\tau_t(x_t) | x_t)] \right\rangle_{q_\pi}, \quad (18)$$

and as $\log P(\bar{r} = 1 | x_0; \pi^0)$ is constant w.r.t. π , the result follows. \square

As an immediate consequence we obtain the direct equivalent for a given stochastic optimal control problem by,

Corollary. *With $\pi^0(\cdot | x) = \mathcal{U}(\cdot)$, where $\mathcal{U}(\cdot)$ is the uniform distribution over \mathcal{U} , the problem in (11) is equivalent to the stochastic optimal control problem.*

One should note, that in general the result requires the set of controls to be such as to allow a uniform distribution to be defined, i.e., either finite or bounded. This is however merely a theoretical consideration, and although we will formally limit ourselves to cases where it is satisfied, it is of little practical consequence.

In general the presented reformulation of the stochastic optimal control problem will remain as intractable as the original formulation. In particular we can see that under the conditions of the corollary the KL divergence reduces directly to the Bellman equation (5) plus a constant. This is a consequence of the fact that minimizing $\text{KL}(q \| p)$, whilst restricting q to be a delta distribution is equivalent to finding the maximum of p . Despite this intractability in the general case we argue that the presented formulation constitutes an interesting starting point for novel approaches to the problem. Both exact, iterative ones, as illustrated in the following section, but also approximate ones, as is the case with the Approximate Inference Control framework of Toussaint [33, 22], to which we relate this result in subsubsection 3.4.1.

3.3 Iterative Solution

From the Bayesian point of view the restriction to delta distributions in Proposition 1 seems rather unnatural and can, as mentioned previously, be seen as the main cause why the KL divergence remains intractable. A relaxation of this restriction, i.e. minimising w.r.t. to an arbitrary distribution $\pi(\cdot | x_t)$, makes as we shall show, the minimization tractable and although it obviously does not lead directly to a optimal policy, we have the following result

Proposition 2. *For any $\pi \neq \pi^0$, $\text{KL}(q_\pi \| p_{\pi^0}) \leq \text{KL}(q_{\pi^0} \| p_{\pi^0})$ implies $\langle \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_\pi} < \langle \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_{\pi^0}}$.*

Proof. Expanding the KL divergences we have

$$\begin{aligned} & \text{KL}(q_\pi(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) + \langle \log P(r_t = 1 | \bar{x}, \bar{u}) \rangle_{q_\pi(\bar{x}, \bar{u})} + \log P(\bar{r} = 1 | x_0; \pi^0) \\ & \leq \text{KL}(q_{\pi^0}(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) + \langle \log P(\bar{r} = 1 | \bar{x}, \bar{u}) \rangle_{q_{\pi^0}(\bar{x}, \bar{u})} + \log P(\bar{r} = 1 | x_0; \pi^0). \end{aligned} \quad (19)$$

Subtracting $\log P(\bar{r} = 1 | x_0; \pi^0)$ on both sides and noting that $\text{KL}(q_{\pi^0}(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) = 0$, we obtain

$$\text{KL}(q_\pi(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) + \langle \log P(\bar{r} = 1 | \bar{x}, \bar{u}) \rangle_{q_\pi(\bar{x}, \bar{u})} \leq \langle \log P(\bar{r} = 1 | \bar{x}, \bar{u}) \rangle_{q_{\pi^0}(\bar{x}, \bar{u})}. \quad (20)$$

Hence, as $\text{KL}(q_\pi(\bar{x}, \bar{u}) \| q_{\pi^0}(\bar{x}, \bar{u})) \geq 0$ with equality iff $\pi = \pi^0$, the result follows. \square

As an immediate consequence, with some initial π^0 , the iteration

$$\pi^{i+1} \leftarrow \underset{\pi}{\operatorname{argmin}} \text{KL}(q_\pi \| p_{\pi^i}), \quad (21)$$

with π an arbitrary² conditional distribution over u , gives rise to a chain of stochastic policies with ever decreasing expected costs. However we note that Proposition 2 has rather weak conditions and we can generalise the iteration as follows³

²n.b. formally certain assumptions have to be made to ensure the support of q_π is a subset of the support of p_{π^i}

³n.b. a more general formulation is possible, which does not require $\pi^i \in \mathcal{P}^i$, however the presented formulation suffices for our purpose

Proposition 3. Let \mathcal{P} be the set over all (stochastic) policies, if $\mathcal{P}^i \subseteq \mathcal{P}$ s.t. $\pi^i \in \mathcal{P}^i$ for all i , then the policies in the sequence generated by

$$\pi^{i+1} \leftarrow \underset{\pi \in \mathcal{P}^i}{\operatorname{argmin}} \operatorname{KL}(q_\pi \| p_{\pi^i}) \quad (22)$$

have non increasing expected costs.

Proof. As $\pi^i \in \mathcal{P}^i$, $\operatorname{KL}(q_{\pi^{i+1}} \| p_{\pi^i}) \leq \operatorname{KL}(q_{\pi^i} \| p_{\pi^i})$ and hence either Proposition 2 applies or $\pi^i = \pi^{i+1}$. \square

Note that this formulation admits (21) as a special case. A further interesting case is what we will refer to as *asynchronous updates*. These are updates of only one time step at each iteration in any particular order, i.e. choose a schedule of time steps $\hat{t}^0, \hat{t}^1, \dots$ and let $\mathcal{P}^i = \{\pi \in \mathcal{P} : \forall t \neq \hat{t}^i, \pi_t = \pi_t^i\}$.

Naturally questions about the behaviour in the limit of iterations covered by Proposition 3 arise. As the expected cost is, under the assumption $\mathcal{C}_t(\cdot) > 0$ (cf. section 2), bounded from below, we have as an immediate consequence

Corollary. Any iteration of the form (22) converges.

This obviously leaves open the more interesting question, if, under what conditions and in what sense the policy converges to an optimal policy. Although it would certainly be desirable to obtain a general answer to this question, currently we are concentrating on the specific cases of (21) and asynchronous updates for which we suggest that under weak conditions convergence to an optimal policy occurs (see Conjecture 5 below).

We will now proceed by first deriving specific updates for the finite horizon case, subsequently extending these to the infinite horizon, discounted cost setting.

3.3.1 Finite Horizon Case

As indicated previously the general minimization of iteration (21) can be performed analytically and here we provide the required derivation for the finite horizon case. To obtain the solution we bring the KL divergence into the recursive form

$$\begin{aligned} \operatorname{KL}(q_{\pi^{i+1}}(\bar{x}, \bar{u}) \| p_{\pi^i}(\bar{x}, \bar{u})) &= \int_{u_0} \pi^{i+1}(u_0 | x_0) \left[\log \frac{\pi^{i+1}(u_0 | x_0)}{\pi^i(u_0 | x_0) P(r_0 | x_0, u_0)} \right. \\ &\quad \left. + \int_{\hat{x}} P(\hat{x} | x_0, u_0) \operatorname{KL}(q_{\pi^{i+1}}(x_{2:T}, u_{1:T} | x_1 = \hat{x}) \| p_{\pi^i}(x_{2:T}, u_{1:T} | x_1 = \hat{x})) \right] \end{aligned} \quad (23)$$

and utilizing the following general result

Lemma 4. Let a, b, c be random variables with joint $P(a, b, c) = P(a)P(b|a)P(c|b, a)$ and \mathcal{P} the set of distributions over a , then

$$P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c} | b)\right\} \propto \underset{q \in \mathcal{P}}{\operatorname{argmin}} \operatorname{KL}(q(a)P(b|a) \| P(a, b | c = \hat{c})) \quad (24)$$

and

$$\int_a P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c} | b)\right\} = \min_{q \in \mathcal{P}} \operatorname{KL}(q(a)P(b|a) \| P(a, b | c = \hat{c})) . \quad (25)$$

Proof. see Appendix A \square

Specifically assume the minimised nested KL divergence for some time step $t+1$ is given by some $\exp\{\bar{\Psi}_{t+1}(x_{t+1})\}$. Using the recursive formulation (23) and applying (24) with $a = u_t | x_t$, $b = x_{t+1}$ and $P(c = \hat{c} | b) = \exp\{\bar{\Psi}_{t+1}(x_{t+1})\} P(r_t | x_t, u_t)$, it is easy to see that the new policy is given by the Boltzmann like distribution,

$$\pi^{i+1}(u_t | x_t) = \exp\{\Psi_t^{i+1}(x_t, u_t) - \bar{\Psi}^{i+1}(x_t)\} , \quad (26)$$

with energy

$$\Psi_t^{i+1}(x_t, u_t) = \log \pi^i(u_t | x_t) + \log P(r_t = 1 | x_t, u_t) + \int_{x_{t+1}} P(x_{t+1} | x_t, u_t) \bar{\Psi}_{t+1}^{i+1}(x_{t+1}) \quad (27)$$

and log partition function

$$\bar{\Psi}_t(x_t) = \log \int_u \exp\{\Psi(x_t, u)\} . \quad (28)$$

Thus we can obtain the result for iteration (21) by applying (27) backwards in time, with $\bar{\Psi}_{T+1}^{i+1} = 0$ as the base case. Similarly asynchronous updates can be obtained by applying (27) only at one time step.

We now turn to the question of the behaviour of these updates in the limit. Let us define the following restricted optimal policy $\bar{\pi}^*$

Definition 1. Let $\mathcal{U}_t^0(x) \subseteq \mathcal{U}$ be the support of $\pi_t^0(\cdot | x)$ and let $\mathcal{U}_t^*(x)$ be the optimal controls at time t in state x . If $\bar{\mathcal{U}}_t^*(x) = \mathcal{U}_t^*(x) \cap \mathcal{U}_t^0(x)$ is not empty, $\bar{\pi}^*(\cdot | x)$ is defined as the uniform distribution over $\bar{\mathcal{U}}_t^*(x)$.

Although we do not have any formal results yet, we suggest the following preliminary conjecture which we aim to complete in the near future

Conjecture 5. *Under weak assumptions, for both (21) and asynchronous updates,*

- π^i converges weakly to $\bar{\pi}^*$
- $\bar{\Psi}_t$ converges pointwise to $-\mathcal{J}_t^* + c_t$, with \mathcal{J}_t^* the optimal value function and c_t a constant

3.3.2 Infinite Horizon Case

We will now consider the discounted infinite horizon setting. We proceed rather informally, but aim in future to formalise this setting as a limit case of the finite horizon setting.

It is sufficient to only consider time stationary policies in this setting [28]. Under such a policy the entire process is time stationary, and, with a slight abuse of notation, we have

$$q_\pi(x_{>1}, u_{>0} | x_0 = \hat{x}) = q_\pi(x_{>2}, u_{>1} | x_1 = \hat{x}) . \quad (29)$$

It is now easy to show that

$$\text{KL}(q_{\pi^{i+1}}(x_{>2}, u_{>1} | x_1 = \hat{x}) \| p_{\pi^i}(x_{>2}, u_{>1} | x_1 = \hat{x})) = \gamma \text{KL}(q_{\pi^{i+1}}(\bar{x}, \bar{u} | x_0 = \hat{x}) \| p_{\pi^i}(\bar{x}, \bar{u} | x_0 = \hat{x})) , \quad (30)$$

which leads to the time stationary analog of (27),

$$\Psi^{i+1}(x, u) = \log \pi^i(u | x) + \log P(r = 1 | x, u) + \gamma \int_y P(y | x, u) \bar{\Psi}^{i+1}(y) . \quad (31)$$

However due to the form of $\bar{\Psi}^{i+1}$, this does not yield Ψ^{i+1} directly. Therefore we propose, in analogy to value iteration, e.g., [28], the update

$$\Psi^{i+1}(x, u) \leftarrow \Psi^i(x, u) - \bar{\Psi}^i(x) + \log P(r = 1 | x, u) + \gamma \int_{x'} P(x' | x, u) \bar{\Psi}^i(x') . \quad (32)$$

which corresponds to the assumption that after one step the old policy $\pi^i = \exp\{\Psi^i(x, u) - \bar{\Psi}^i(x)\}$ is followed. Although this update does not correspond to the iteration of (21), it can be constructed from a specific schedule of asynchronous updates. Specifically consider the schedule given with $\hat{t}^{j,k}$, where for each $j = 1, 2, \dots$ updates are performed at $k = j, j-1, j-2, \dots, 0$. It is easy to see that after each update $\hat{t}^{j,0}$, the first step policy equals π_0^i , the policy obtained from (32). Hence as this iteration falls into the class of Proposition 3 we immediately obtain the guarantee of non increasing expected costs and convergence. Furthermore we anticipate that the schedule will satisfy the weak conditions of Conjecture 5 and its convergence to an optimal policy will directly follow.

3.4 Relation to Previous Work

In the following we will relate the presented work in greater detail to three recent developments in the field. However we note that attempts to relate stochastic optimal control to inference have along history, in part motivated by the exact duality for the *linear-quadratic-gaussian* (LQG) case discovered by Kalmann [27]. In general the idea of replacing costs, utilities or rewards by an auxiliary binary random variable has a long history [7, 25, 8]. Shachter & Peot [26] even mention work by Raiffa (1969) and von Neumann & Morgenstern (1947) in this context. Although approaches have varied between using the interpretation of cost as energy, together with the typical identification of energy with negative log probability, as has been done here, and choosing probabilities which are proportional to the reward or utility.

3.4.1 Approximate Inference Control

As the *approximate inference control* (AICO) framework was discussed in detail in the 1st Year Report we will refrain from a full description here. It suffices to recall that AICO is formulated within the model described in subsection 3.1 and aims to find an approximation to p_{π^0} by a message passing approach similar to Expectation Propagation [16]. Although the original work [33] observed a close relation of the messages in the LQG case to the classical Riccati [27] equations, no claims were made regarding stochastic optimality and it was suggested to choose the *maximum a posteriori* (MAP) controls. With the results presented in subsection 3.2 the relation of AICO to stochastic optimal control can now be clarified. Specifically a possible interpretation for AICO is to see it as finding an approximation to p_{π^0} , such as to make the KL divergence of Proposition 1 tractable. However, even under this interpretation we note that the result of the minimization of the KL divergence, even under a Gaussian approximation to p_{π^0} , are not the MAP control, rather one should solve the Ricatti equation arising from the approximation.

3.4.2 Path Integral and KL control

In recent years several groups were independently able to show that for a restricted class of stochastic optimal control problems the minimized Bellman equation (5) becomes linear and the problem admits a solution in closed form [18, 13, 31, 12]. These linear Bellmann equations can be seen as a KL divergence [12], leading to a close relation to the formulation in Proposition 1. We will demonstrate this close relation in the discrete time case, leaving the continuous time case for future consideration as we have not yet developed it in our framework.

Let us briefly recall the KL control framework of Kappen et.al. [12], the alternative formulations of Todorov [30, 31] being equivalent. Choose some free dynamics $\nu_0(x_{t+1}|x_t)$ and let the cost be given as

$$\mathcal{C}(\bar{x}) = \ell(\bar{x}) + \sum \log \frac{\nu(\bar{x})}{\nu_0(\bar{x})} \quad (33)$$

where $\nu(x_{t+1}|x_t)$ is the controlled process under some policy. Then

$$\langle \mathcal{C}(\bar{x}) \rangle_\nu = \text{KL}(\nu(\bar{x}) \| \nu_0(\bar{x}) \exp\{-\ell(\bar{x})\}) \quad (34)$$

which is minimised w.r.t. ν by

$$\nu(x_{1..T}|x_0) = \frac{1}{Z(x_0)} \exp\{-\ell(x_{1..T})\} \nu_0(x_{1..T}|x_0) \quad (35)$$

and one concludes that the optimal control is given by $\nu(x_{t+1}|x_t)$, where presumably the implied meaning is that $\nu(x_{t+1}|x_t)$ is the trajectory distribution under the optimal policy.

Although (35) gives a process which minimises (34), it is not obvious how to compute actual controls from this process. Specifically when given a model of the dynamics, $P(x_{t+1}|x_t, u_t)$, and having chosen some ν_0 , a non trivial, yet implicitly made, assumption is that

$$\exists \pi, \quad \text{s.t.} \quad \nu(x_{t+1}|x_t) = \int_{u_t} P(x_{t+1}|x_t, u_t) \pi(u_t|x_t) . \quad (36)$$

In fact in general such a π will not exist. This is made very explicit for the discrete MDP case in [30], where it is acknowledged that the method is only applicable if the dynamics are fully controllable, i.e., $P(x_{t+1}|x_t, u_t)$ can be brought into any arbitrary form by the controls. Although in the same paper it is suggested that solutions to classical problems can be obtained by *continuous embedding* of the discrete MDP, such an approach has several drawbacks. For one it requires solving a continuous problem even for cases which could have been otherwise represented in tabular form, but more importantly such an approach is obviously not applicable to problems which already have continuous state or action spaces. In the latter case Kappen et.al. claim (cf. section 4 of [12]) that the KL control approach is applicable if the problem is of the following form

$$\begin{aligned} x_{t+1} &= \mathcal{F}(x_t) + \mathbf{B}(x_t)(u_t + \xi), \quad \xi \sim \mathcal{N}(0, \mathbf{Q}), \\ \mathcal{C}_t(x_t, u_t) &= \ell(x_t) + u_t^T \mathbf{H} u_t, \end{aligned} \quad (37)$$

with \mathcal{F}, \mathbf{B} and ℓ having arbitrary form, but \mathbf{H}, \mathbf{Q} are such that $\mathbf{H}^{-1} \propto \mathbf{Q}$. We dispute this claim, showing that, in the discrete time case, (36) is not fulfilled and that correcting the problem leads to an equivalent of Proposition 1.

It will be sufficient to consider the simplest possible case of a one dimensional, one time step LQG problem. Let

$$P(x_{t+1}|x_t, u_t) = \mathcal{N}(x_{t+1}|x_t + u_t; \Sigma) \quad (38)$$

and

$$\mathcal{C}_t(x_t, u_t) = x_t R x_t + u_t \Sigma^{-1} u_t. \quad (39)$$

The claim made by Kappen et.al. is, that for $\nu_0 = P(x_{t+1}|x_t, u_t = 0)$, the KL formulation is equivalent to the corresponding stochastic optimal control problem. Or more specifically that

$$\nu(x_1|x_0) \propto P(x_1|x_0, u_0 = 0) \exp\{-x_1 R x_1\} = \mathcal{N}(x_1|x_0; \Sigma^{-1} + R) \quad (40)$$

gives the optimal controls, hence (36) should hold. In particular, as we know the LQG problem has a unique deterministic stochastic optimal control solution [27], there should be a deterministic π s.t. (36) holds. But notice that we can not influence the variance of $P(x_{t+1}|x_t, u_t)$ by specific choices of a deterministic π , hence (36) does not hold. Specifically ν is *not* the trajectory distribution under the optimal policy. In fact there may not even be a stochastic policy s.t. (36) holds. Consider the case when the cost 'variance' R^{-1} is smaller than the variance of the noise, Σ . Then $\nu(x_1|x_0 = 0)$ will have variance smaller than Σ . But even though with a stochastic policy the variance of the marginal process can increase, it can not decrease.

The question now arises what controls should we choose? A principled choice would be to choose π to minimise a KL divergence. The first intuition would be to take

$$\operatorname{argmin}_{\tau} \text{KL}(P(x_{t+1}|x_t, u_t = \tau(x_t)) \| \nu(x_{t+1}|x_t)) \quad (41)$$

However noting that

$$\nu(x_{t+1}|x_t) = \frac{1}{Z(x_t)} \nu_0(x_{t+1}|x_t) Z(x_{t+1}) \quad (42)$$

$$= \frac{1}{Z(x_t)} \nu_0(x_{t+1}|x_t) \langle \exp\{-\mathcal{C}(x_{k+1:K})\} \rangle_{\nu_0(x_{k+1:K}|x_{t+1})}, \quad (43)$$

the KL divergence can be written as

$$\text{KL}(P(x_{t+1}|x_t, u_t = \tau(x_t)) \| \nu_0(x_{t+1}|x_t)) + \langle \log Z(x_{t+1}) \rangle_{P(x_{t+1}|x_t, u_t = \tau(x_t))} - \log Z(x_t). \quad (44)$$

This is the correct expression for the expected cost, if $\log Z$ is the value function, however the latter is only the case if the normalized form of the KL divergence in (34) becomes zero at the minimum. Here we are specifically assuming this not to be the case, implying this formulation does not lead to stochastic optimal controls and we are therefore compelled to take

$$\operatorname{argmin}_{\pi} \text{KL}(q_{\pi}(\bar{x}) \| \nu(\bar{x})) \quad (45)$$

This is very similar to the KL divergence in Proposition 1. In fact, under the conditions of the corollary to Proposition 1 and if the problem is of the form in (37), we can write

$$p_{\pi^0}(\bar{x}, \bar{u}) = \nu(\bar{x}) \prod \exp\{(x_{t+1} - x_t - \mathcal{F}(x_t)\Sigma^{-1}u_t - \frac{1}{2}u_t H^{-1}u_t)\} \quad (46)$$

and the KL divergence of Proposition 1 can alternatively be written as

$$\text{KL}(q_{\pi} \| p_{\pi^0}) = \text{KL}(q_{\pi}(\bar{x}) \| \nu(\bar{x})) - \left\langle \sum (x_{t+1} - x_t - f(x_t)\Sigma^{-1}u_t - \frac{1}{2}u_t H^{-1}u_t) \right\rangle_{q_{\pi}(\bar{x}, \bar{u})} . \quad (47)$$

Furthermore as for a deterministic policy, i.e. $\pi(u_t | x_t) = \delta_{u_t = \tau(x_t)}$,

$$\langle (x_{t+1} - x_t - f(x_t)) \rangle_{q_{\pi}} = \langle u_t \rangle_{q_{\pi}} = \tau(x_t) , \quad (48)$$

we can see that the second term is zero under the condition $H^{-1} = 2\Sigma^{-1}$, i.e. under the conditions required by Kappen et.al., and (45) is equivalent to the formulation in Proposition 1.

3.4.3 Expectation Maximization Approaches

Several suggestions for mapping the stochastic optimal control problem onto a maximum likelihood problem and using *Expectation Maximization* (EM) have been recently made in the literature [34, 2]. Going further back the probability matching approach [8, 24] is also closely related to expectation maximization procedures.

As one may suspect when considering (21) our approach has a close relation to the free energy view of EM [19, 4]. In this view, EM alternates between minimizing $\text{KL}(q(z) \| P(z|y; \theta))$ w.r.t. q , where z, y are the latent and observed variables and θ the parameters, and maximizing the free energy, defined as

$$\mathcal{L}(q, \theta) = \int_z q(z) \log \frac{P(z, y; \theta)}{q(z)} \quad (49)$$

w.r.t. θ . In our case z and y correspond to \bar{x}, \bar{u} and \bar{r} , while θ corresponds to π . It is easy to see that (21), or (22) for that matter, correspond to a generalized E-Step. The *generalized* indicates that only a partial step is performed, i.e., we are only lowering, rather than minimizing, $\text{KL}(q(z) \| p(z|y; \theta))$ w.r.t. q . Furthermore the choice of π^{i+1} corresponds to a generalized M-Step, as

$$\mathcal{L}(q_{\pi^{i+1}}, \pi = \pi^{i+1}) = \int_{\bar{x}, \bar{u}} q_{\pi^{i+1}} \log P(\bar{r} = 1 | \bar{x}, \bar{u}) \quad (50)$$

$$\geq \int_{\bar{x}, \bar{u}} q_{\pi^{i+1}} \log P(\bar{r} = 1 | \bar{x}, \bar{u}) - \text{KL}(q_{\pi^{i+1}} \| q_{\pi^i}) \quad (51)$$

$$= \mathcal{L}(q_{\pi^{i+1}}, \pi = \pi^i) . \quad (52)$$

Hence we conclude that our method corresponds to an generalized EM algorithm.

Although we have shown that one can interpret the proposed approach in terms of EM we emphasise that it differs significantly from the applications of EM in previous work. For one we note that it is not our aim to find the maximum likelihood policy and in fact as we are using a generalized E-Step we lose the guarantee of convergence to a local maximum of the likelihood. In general maximizing the marginal log likelihood, the objective of EM, would not be desirable in our model, as despite the fact that for a given state and control trajectory, the classical cost and the task likelihood are directly related by

$$\mathcal{C}(\bar{x}, \bar{u}) = -\log P(\bar{r} = 1 | \bar{x}, \bar{u}) , \quad (53)$$

no such direct equality relation for the marginal likelihood can be obtained. Although using Jensen's inequality we may obtain

$$\langle \mathcal{C}(\bar{x}, \bar{u}) \rangle_{q_{\pi}(\bar{x}, \bar{u})} \leq -\log P(\bar{r} = 1) , \quad (54)$$

this bound, which has also been previously observed in [33], is not necessarily tight and hence the optimal stochastic optimal solution does not necessarily coincide with the maximum likelihood solution.

A more fundamental difference is that we can avoid finding an explicit representation for q_π . In both the approaches of [34] and [2] calculating q_π explicitly is a major computational step and presents a problem if these methods were to be applied to the continuous setting where q_π may not be analytically tractable.

Finally we anticipate that Conjecture 5 will hold, giving a guarantee of convergence to an optimal policy which other EM methods can not provide.

3.5 Conclusion

The contribution of this section is a novel interpretation of the stochastic optimal control problem as an approximate inference problem and the derivation of a iterative solution to the control problem based on this new interpretation. The proposed approach has also been shown to have interesting links to other current research directions in the field. In particular we demonstrate that the approach can be understood to underlie both the approximate inference control framework and, in the time discrete setting, the KL control framework. This theoretical work is intended to provide the foundation for the remainder of this paper and future work.

4 Reinforcement Learning

So far we have assumed the transition model and cost function are readily available. We now turn to the reinforcement learning setting [11, 29, 28], where one aims to learn a good policy only given samples from the transition probability and associated incurred costs.

We will demonstrate how the theoretical results previously derived can be applied to such problems yielding algorithms which are both *model free* and *off policy*. Model free indicates that the algorithm does not construct an explicit representation of the transition probability and cost function but rather directly learns a representation of the optimal policy. Off policy on the other hand means that the optimal policy can be learnt from samples collected under a different, often sub-optimal, policy.

We will proceed by first deriving a tabular algorithm which is applicable for problems with small, finite, state and control spaces, before subsequently extending it to problems with continuous state and control spaces by using approximate parametric representations. Both algorithms are applied to classical problems in the field.

4.1 Finite Problems

Let us consider problems in the infinite horizon discounted cost setting and recall that the update function for Ψ suggested in subsection 3.3 for this case was

$$\Psi(x, u) \leftarrow \Psi(x, u) - \bar{\Psi}(x) + \log P(r = 1|x, u) + \gamma \int_{x'} P(x'|x, u) \bar{\Psi}(x') . \quad (55)$$

For any given x, u this update can be written as an expectation w.r.t. the transition probability $P(y|x, u)$, and hence may be approximated from a set of sampled transitions. In particular given a single sample (x, u, ℓ, y) of a transition from x to y under control u incurring cost⁴ ℓ we may perform the approximate update

$$\Psi(x, u) \leftarrow \Psi(x, u) + [\gamma \bar{\Psi}(y) - \bar{\Psi}(x) - \ell] . \quad (56)$$

Given a stream of samples $x_0, u_0, \ell_0, x_1, u_1, \ell_1, \dots$ we can then apply such an update for each tuple $(x_t, u_t, \ell_t, x_{t+1})$ individually. Without a particular justification we furthermore can introduce a decaying learning rate parameter, similar to other reinforcement learning algorithms, in order to damp these updates. In practise however we did not find such a learning rate to improve results significantly. We call the resulting algorithm *Ψ -learning*. As indicated previously it is model free and can be employed for off policy learning.

⁴n.b. we assume we observe the cost, i.e., $\ell = -\log P(r = 1|x, u)$.

4.1.1 Relation to Classical Algorithms

Before proceeding let us highlight certain similarities and differences between Ψ -learning and two classical algorithms, Q -learning and TD(0) [28].

As the name indicates, Q -learning learns the state-action value function (cf. Equation (4)). We note that Ψ has certain similarities to a Q function, in the sense that a higher value of Ψ for a certain control in a given state indicates that the control is 'better'. In fact for the optimal controls the Q function and Ψ converge to the same value⁵. However unlike the Q function, which also converges to the expected cost for the sub-optimal controls, Ψ goes to $-\infty$ for sub-optimal actions. A potentially more insightfull difference between the two algorithm is nature of updates employed. The Q -learning algorithms uses updates of the form

$$Q(x, u) \leftarrow Q(x, u) + \alpha \left[\ell + \gamma \max_{u'} Q(y, u') - Q(x, u) \right], \quad (57)$$

where α is a learning rate. Note that it will employ only information from one current control and the best, according to current knowledge, future control. The Ψ -learning algorithm on the other hand uses $\bar{\Psi}$ which in some sense averages over information about the future according the current belief about the control distribution, rather than using single Ψ values.

A connection to the TD(0) algorithm which learns a value function is given by the form of the update. The TD(0) update has the form

$$\mathcal{J}(x) = \mathcal{J}(x) + \alpha [\ell + \gamma \mathcal{J}(y) - \mathcal{J}(x)] \quad (58)$$

with α again a learning rate. We observe that as by Conjecture 5, $\bar{\Psi}$ converges, up to a additive constant, to the value function of the optimal policy, the Ψ -learning update converges towards the TD(0) update for samples generated under the optimal policy. The emphasise is on, *convergence* to the TD(0) update, in general it will not correspond to an TD(0) update. In particular a important differences between the two algorithms is that TD(0) is a on-policy method, that is it learns the value function of the policy used to generate samples, while the proposed Ψ -learning is off-policy.

4.1.2 Results

Problems with finite state and action spaces allow Ψ to be represented directly in tabular form. We evaluated such a tabular Ψ -learning algorithm on the grid world domain [28]. Specifically we used the following task formulation. The state space is given by a $N \times N$ grid with some states occupied by obstacles. The controls allow the agent to transition to any neighbouring state not occupied by an obstacle or to remain at the current state. A transition to a neighbouring state succeeds with probability 0.8, with the agent remaining at the current location in case of failure. Choosing to remain in the current state succeeds with probability 1. Additionally a set $\mathcal{A} \subseteq \mathcal{X}$ of absorbing target states, i.e., $1 = P(x_{t+1} \in \mathcal{A} | x_t \in \mathcal{A}, u \in \mathcal{U})$, is defined. In every time step a cost of 1 is incurred if the agent is in any state which is not a target state, while at a target state no cost is incurred, i.e., $C(x, u) = \delta_{x \notin \mathcal{A}}$ with δ the Kronecker delta. The cost was not discounted, i.e., $\gamma = 1$.

We used tabular Q -learning, e.g., [28], as a baseline. Both algorithms were run with controls sampled from an uninformed policy, i.e. a uniform distribution over the controls available at a state. Once a target state was reached, or if the target wasn't reached within 100 steps, the state was reset randomly. The learning rate for Q -learning decayed as $\alpha = c/(c + t)$ with t the number of transitions sampled and c a constant which was optimised manually.

Representative results for a single instance of the general task are given in figure 2. We report the approximation error

$$e_{\mathcal{J}} = \frac{\max_x |\mathcal{J}(x) - \hat{\mathcal{J}}(x)|}{\max_x \mathcal{J}(x)} \quad (59)$$

⁵n.b. at the moment this is conjecture, as it is a consequence of Conjecture 5

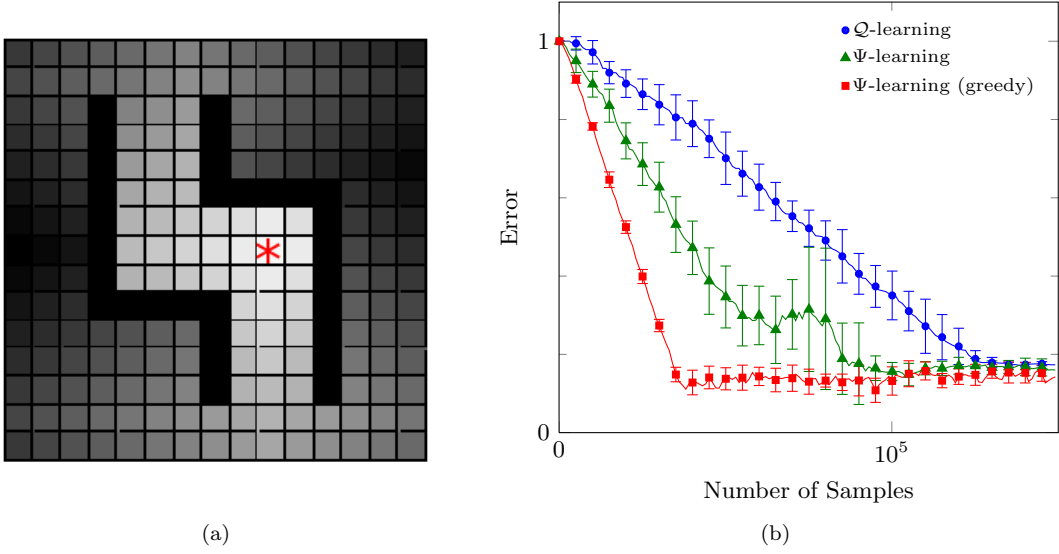


Figure 2: Results for tabular Ψ -learning on an example grid world problem. **(a)** the optimal value function (white low expected cost - black high expected cost) of the problem. Obstacles are black and the target state is indicated by *. **(b)** Evolution of the mean error in (59) averaged over 10 trials for each of the algorithms. Error bars indicate the standard deviation.

between the true value function \mathcal{J} , obtained by value iteration, and its estimate \hat{J} , given by $\bar{\Psi}$ and $\max_u Q(x, u)$ respectively. Both algorithm achieved the same error at convergence. However Ψ -learning consistently outperformed Q -learning in terms of the number of samples required to convergence. We additionally considered a greedy variant of Ψ -learning where the controls are sampled from the policy given by the current Ψ , i.e. $\pi(u|x) = \exp\{\Psi(x, u) - \bar{\Psi}(x)\}$. As expected we found that the greedy version greatly outperformed sampling using an uninformed policy.

4.2 Continuous problems

For continuous control problems, i.e. those with infinite state or controls sets, storing Ψ in tabular form clearly becomes impossible and even for discrete problems it may be impracticable due to the size of the table required. In such cases, it is common to resort to parametric representations [29], and here we follow such an approach to extend Ψ -learning to continuous problems. Although we will concentrate on the continuous case, we note that the proposed approach could also be employed for large discrete problems.

4.2.1 The LS Ψ algorithm

Similar to numerous previous approaches [5, 20, 28, 29] we used a linear basis function model, to approximate Ψ , i.e.,

$$\Psi(x, u) \approx \tilde{\Psi}(x, u, \mathbf{w}) = \sum_{i=0}^M w_i \phi_i(x, u) \quad (60)$$

where $\phi_i : \mathcal{X} \times \mathcal{U} \rightarrow \mathbb{R}$ are a set of given basis functions and $\mathbf{w} = (w_1, \dots, w_M)$ is the vector of parameters we learn. For such an approximation and given set of samples $(x_{1 \dots K}, u_{1 \dots K}, \ell_{1 \dots K}, y_{1 \dots K})$, the Ψ -learning update (56) can be written in matrix notation as

$$\Phi \mathbf{w}^{i+1} = \Phi \mathbf{w}^i + \mathbf{z}, \quad (61)$$

where Φ is the $K \times M$ matrix with entries $\Phi_{i,j} = \phi_i(x_j, u_j)$ and \mathbf{z} is the vector with elements

$$\mathbf{z}_k = \gamma \bar{\Psi}(y_k) - \ell_k - \bar{\Psi}(x_k) . \quad (62)$$

From this we can obtain

$$\mathbf{w}^{i+1} - \mathbf{w}^i = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{z} , \quad (63)$$

which suggests the update rule

$$\mathbf{w} \leftarrow \mathbf{w} + (\Phi^T \Phi)^{-1} \Phi^T \mathbf{z} . \quad (64)$$

This is equivalent to computing the Ψ -learning update of (56) for the current approximation and projecting the result onto the space spanned by the basis functions in the least squares sense, an approach which has seen repeated use in reinforcement learning [14, 29]. We call the algorithm resulting algorithm, which, as tabular Ψ -learning, is a model free, off-policy method, *Least Squares Ψ -learning* (LS Ψ).

The choice of basis functions for LS Ψ is somewhat complicated by the need to evaluate the log partition function of the policy $\tilde{\Psi}$, i.e. $\log \int_u \exp\{\tilde{\Psi}(x, u)\}$, when forming the vector \mathbf{z} . In cases where \mathcal{U} is a finite set, arbitrary basis functions can be chosen as the integral reduces to a finite sum. However for problems with infinite control spaces one needs to ensure the bases are chosen such that the arising integral is analytical tractable, i.e. the partition function of the stochastic policy can be evaluated. One class of basis sets for which this is the case, are those for which $\tilde{\Psi}(x, u, \mathbf{w})$ has the form

$$\tilde{\Psi}(x, u, \mathbf{w}) = -\frac{1}{2} u^T \mathbf{K}(x, \mathbf{w}) u + u^T \mathbf{k}(x, \mathbf{w}) + k(x, \mathbf{w}) \quad (65)$$

where $\mathbf{K}(x, \mathbf{w})$ is a positive definite matrix. For such a set the integral is of the Gaussian form and the closed form solution

$$\log \int_u \exp\{\tilde{\Psi}\} = -\log |\mathbf{K}| - \frac{1}{2} \mathbf{k}' \mathbf{K}^{-1} \mathbf{k} + k + \text{constant} \quad (66)$$

is obtained. Obviously the implication of such a basis set is that the policies are restricted to conditional Gaussian distributions. Specifically the policy is given by

$$\pi(u|x, \mathbf{w}) = \mathcal{N}(u | \mathbf{K}^{-1} \mathbf{k}, \mathbf{K}^{-1}) . \quad (67)$$

Such Gaussian policies are commonly employed in the continuous reinforcement learning setting, e.g., [6, 20], and we emphasise that, as the state dependent part of the basis is largely unrestricted, this general class of basis sets does not seem unreasonably restrictive.

4.2.2 Results

We demonstrate the applicability of LS Ψ on a pole on cart task [28], which has been repeatedly used as a benchmark in reinforcement learning [21, 23]. The task consists of balancing a inverted pendulum mounted on a cart by exerting forces on the latter. The state space is given by $\mathbf{x} = (x, \dot{x}, \theta, \dot{\theta})$, with x the position of the cart, θ the pendulums angular deviation from the upright postion and $\dot{x}, \dot{\theta}$ their respective temporal derivatives. Following [20] we use a form of the dynamics linearised around the zero state. The approximate dynamics are given by $P(x_{t+1}|x_t, u_t) = \mathcal{N}(x_{t+1} | \mathbf{A}x_t + \mathbf{b}u_t, \Sigma)$, where

$$\mathbf{A} = \begin{bmatrix} 1 & \tau & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tau \\ 0 & 0 & \nu\tau & 1 \end{bmatrix} , \quad \mathbf{b} = \begin{bmatrix} 0 \\ \tau \\ 0 \\ \nu\tau/g \end{bmatrix} \quad (68)$$

and $\tau = 1/60s$, $\nu = 13.2s^{-2}$, $g = 9.8ms^2$, $\Sigma = \text{diag}(0.001, 0.001, 0.001, 0.001)$. The cost is given by $\mathcal{C}(x, u) = \mathbf{x}^T \mathbf{Q} \mathbf{x} + u \mathbf{R} u$, with $\mathbf{Q} = \text{diag}(1.25, 1, 12, 0.25)$ and $\mathbf{R} = 0.01$, and was unlike in [20] not discounted, i.e. $\gamma = 1$.

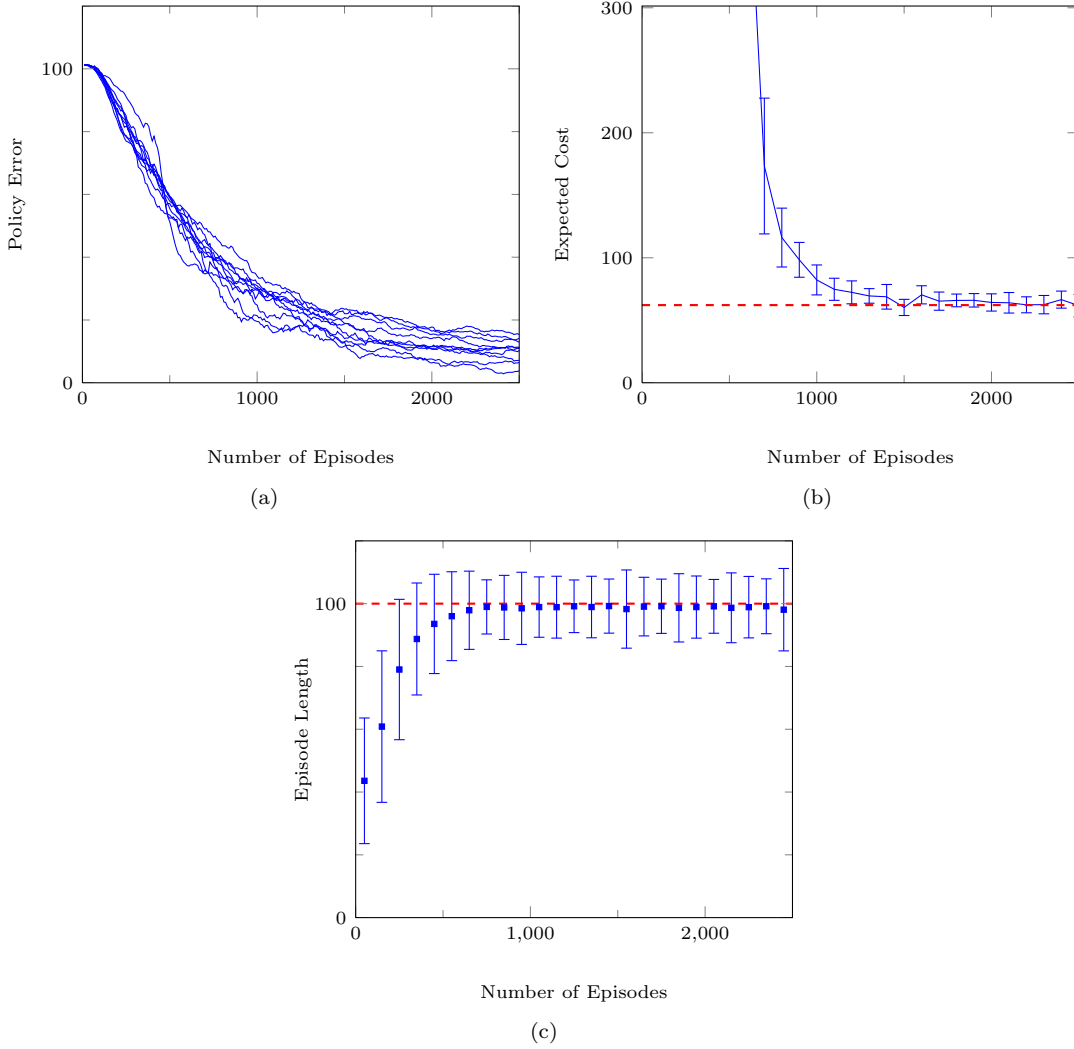


Figure 3: Result for LSΨ for the cart on pole task. **(a)** Evolution of the error in the policy defined as the L_2 norm of the difference between learned and optimal gains for 10 random trials. **(b)** The evolution of the expected cost averaged over the 10 trials. The dashed line indicates the expected cost of the optimal policy. Error bars indicate standard deviation. **(c)** Average length of episodes in the 10 trials, averaged over blocks of 100 episodes.

As the problem is LQG a set of polynomial basis functions is sufficiently rich to capture it, and we applied LS Φ with bases

$$\{u^2, ux, u\dot{x}, u\theta, u\dot{\theta}, x^2, x\dot{x}, x\theta, x\dot{\theta}, \dot{x}^2, \dot{x}\theta, \dot{x}\dot{\theta}, \theta^2, \theta\dot{\theta}, \dot{\theta}^2\}.$$

This set is of the form required by (65), if w_1 is negative. Although we employed no formal means to ensure this, empirically we found that if initialised to a negative value, w_1 would remain negative. Specifically we used the initialisation $\mathbf{w} = (-0.1, 0, \dots, 0)$ corresponding to an initial uniformed policy, i.e. a zero mean Gaussian with large variance. Using a random initialisation, whilst ensuring w_1 was negative did not affect the results significantly, although convergence times increased if the initial policy was far from the optimum and had a low variance. It is worth noting that the initial policy did not asymptotically stabilise the system, this is in contrast to [20, 23]⁶.

We applied LS Ψ following an episodic sampling procedure. Starting from a start state, drawn from $\mathcal{N}(\mathbf{x}_0|0, \Sigma_0)$, with $\Sigma_0 = \text{diag}(0.5, 0, 0.1, 0)$, a state, control & cost trajectory was sampled according to the transition probability and cost function. The required controls were sampled according to the policy arising from the current \mathbf{w} , with a fixed baseline added to the variance. The latter proved necessary as otherwise the updates tended to become numerically unstable once the policy began to converge. A trajectory was terminated when it left the acceptable region given by

$$-\pi/6 \leq \theta \leq \pi/6 \quad \text{and} \quad -1.5m \leq x \leq 1.5m, \quad (69)$$

as in [20], or after 100 time steps. We updated \mathbf{w} after every 10 episodes.

As the problem is LQG, the optimal policy is linear and can be computed directly. We can therefore asses the behaviour of LS Ψ directly, by measuring the error in the policy approximation during learning process. The results in figure 3(a), where we plot the policy error defined as the L_2 norm of the difference between the optimal gains and the LS Ψ estimate, demonstrate that LS Ψ can successfully find near optimal gains. As a, in the literature, more commonly reported metric of the quality of an RL algorithm is the expected cost under the policy it learns. In figure 3(b) we therefore plot the evolution of expected costs. Note that as the expected cost under certain policies for this problem is not finite, we plot a Monte Carlo estimate calculated from a set of 100 trajectories with 200 steps each⁷. As a reference we also plot the expected cost under the optimal policy. This data confirms the results of the policy error analysis, i.e. that LS Ψ converges towards a near optimal policy. As an aside we note that these results are comparable in terms of the convergence time to the best performing methods in [21] were the same problem was used for evaluation. Unfortunately we were, so far, not able to directly reproduce these results in order to obtain a direct comparison. The similar convergence times are in particular surprising as LS Ψ started with a substantially worse initial policy. While [21] seem to have constrained the initial policies to be stable, the initial LS Ψ policy was unstable. This initial instability of the controlled system is illustrated in figure 3(c), where we plot the average length of the episodes used during learning. As can be seen the episodes under the initial policy are significantly shorter than the maximum length, indicating that the constraints in (69) are frequently violated. However after about 600-700 episodes a stabilising policy is learnt.

4.3 Conclusion

The contribution of this section is a novel type of reinforcement learning algorithms, which we obtained by direct application of the theoretical insights of section 3. We were able to demonstrate that the proposed algorithm successfully solves classical problems. However we acknowledge that the performance compared to the state of the art remains to be investigated and we refrain from a full discussion until such data has been obtained.

⁶[20] did not require the initial policy to be stable, however a discounted cost was used and the initial policy was restricted to give $\gamma^{-2} > \text{eig}(\mathbf{A} - \mathbf{bK})$ with K the control gains, i.e. the policy had give rise to a well defined value function

⁷n.b. for the evaluation we did not apply constraints (69)

References

- [1] Mohammad G. Azar and B. Kappen. Dynamic policy programming. ArXiv e-prints 1004.2027, 2010.
- [2] David Barber and Tom Furnston. Solving deterministic policy (PO)MDPs using expectation-maximisation and antifreeze. In *European Conference on Machine Learning (LEMIR workshop)*, 2009.
- [3] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA., 1995.
- [4] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
- [5] J. A. Boyan. technical update: Least-squares temporal difference learning. *Machine Learning*, 49:233–246, 2002.
- [6] Steven J. Bradtke, B. E Ydstie, and Andrew G. Barto. Adaptive linear quadratic control using policy iteration. Technical report, Amherst, MA, USA, 1994.
- [7] G.F. Cooper. A method for using belief networks as influence diagrams. In *Proc. of the Fourth Workshop on Uncertainty in Artificial Intelligence*, pages 55–63, 1988.
- [8] P. Dayan and G. E. Hinton. Using expectation maximization for reinforcement learning. *Neural Computation*, 9:271–278, 1997.
- [9] Jörn Diedrichsen, Reza Shadmehr, and Richard B. Ivry. The coordination of movement: optimal feedback control and beyond. *Trends in Cognitive Sciences*, 14:31–39, 2009.
- [10] D Jacobson and D Mayne. *Differential Dynamic Programming*. Elsevier, 1970.
- [11] L.P. Kaelbling, M.L. Littman, and A.W. Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.
- [12] B. Kappen., V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. ArXiv e-prints 0901.0633, 2009.
- [13] H J Kappen. Path integrals and symmetry breaking for optimal control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11011, 2005.
- [14] Michail G. Lagoudakis and Roland Parr. Least-Squares policy iteration. *Journal of Machine Learning Research*, 4:1107–1149, 2003.
- [15] Weiwei Li and Emanuel Todorov. An iterative optimal control and estimation design for nonlinear stochastic system. In *Proc. of the 45th IEEE Conference on Decision and Control*, 2006.
- [16] Thomas Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, 2001.
- [17] Djordje Mitrovic, Sho Nagashima, Stefan Klanke, Takamitsu Matsubara, and Sethu Vijayakumar. Optimal feedback control for anthropomorphic manipulators. In *Proc. IEEE International Conference on Robotics and Automation (ICRA 2010)*, 2010.
- [18] Sanjoy K. Mitter and Nigel J. Newton. A variational approach to nonlinear estimation. *SIAM J. Control Optim.*, 42(5):1813–1833, 2003.
- [19] Radford Neal and Geoffrey Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, 1999.
- [20] J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *ieee-ras international conference on humanoid robots (humanoids2003)*, 2003.

- [21] Jan Peters and Stefan Schaal. Policy gradient methods for robotics. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 2219–2225, oct. 2006.
- [22] Konrad Rawlik and Marc Toussaint. Approximate inference control. *Journal of Machine Learning Research*, 2010. submitted.
- [23] M. Riedmiller, J. Peters, and S. Schaal. Evaluation of policy gradient methods and variants on the cart-pole benchmark. In *Approximate Dynamic Programming and Reinforcement Learning, 2007. ADPRL 2007. IEEE International Symposium on*, pages 254–261, 1-5 2007.
- [24] Philip N. Sabes and Michael I. Jordan. Reinforcement learning by probability matching. In *Advances in Neural Information Processing Systems*, volume 8, 1996.
- [25] R. D. Shachter. Probabilistic inference and influence diagrams. *Operations Research*, 36:589–605, 1988.
- [26] R.D. Shachter and Peot. Decision making using probabilistic inference methods. In *Proc. of the Eighth Conf. on Uncertainty in Artificial Intelligence*, pages 276–283, 1992.
- [27] Robert F. Stengel. *Optimal Control and Estimation*. Dover Publications, 1986.
- [28] R.S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, 1998.
- [29] Csaba Szepesvri. Reinforcement learning algorithms for mdps – a survey. Technical report, University of Alberta, 2009.
- [30] Emanuel Todorov. Linearly-solvable markov decision problems. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1369–1376. MIT Press, Cambridge, MA, 2007.
- [31] Emanuel Todorov. Efficient computation of optimal actions. *PNAS*, 106:11478–11483, 2009.
- [32] Emanuel Todorov and Michael Jordan. Optimal feedback control as a theory of motor coordination. *Nature Neuroscience*, 5(11):1226–1235, 2002.
- [33] Marc Toussaint. Robot trajectory optimization using approximate inference. In *Proc. of the 26 th International Conference on Machine Learning (ICML 2009)*, 2009.
- [34] Marc Toussaint, Stefan Harmeling, and Amos Storkey. Probabilistic inference for solving (PO)MDPs. Technical Report EDI-INF-RR-0934, University of Edinburgh, School of Informatics, 2006.
- [35] Marc Toussaint, Nils Plath, Tobias Lang, and Nikolay Jetchev. Integrated motor control, planning, grasping and high-level reasoning in a blocks world using probabilistic inference. In *Proc. IEEE International Conference on Robotics and Automation (ICRA 2010)*, 2010.

Appendices

A Supplementary proofs

Lemma (Lemma 4 in subsubsection 3.3.1). *Let a, b, c be random variables with joint $P(a, b, c) = P(a)P(b|a)P(c|b, a)$ and \mathcal{P} the set of distributions over a , then*

$$P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c}|b)\right\} \propto \underset{q \in \mathcal{P}}{\operatorname{argmin}} \operatorname{KL}(q(a)P(b|a) \| P(a, b|c = \hat{c})) \quad (70)$$

and

$$\int_a P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c}|b)\right\} = \min_{q \in \mathcal{P}} \operatorname{KL}(q(a)P(b|a) \| P(a, b|c = \hat{c})) . \quad (71)$$

Proof. We form the Lagrangian

$$\mathcal{L} = \operatorname{KL}(q(a)P(b|a) \| P(a, b|c = \hat{c})) + \lambda \left[\int_a q(a) - 1 \right] \quad (72)$$

$$\cong \int_{a,b} q(a)P(b|a) \log \frac{q(a)P(b|a)}{P(a)P(b|a)P(c = \hat{c}|b)} + \lambda \left[\int_a q(a) - 1 \right] \quad (73)$$

$$= \int_a q(a) \log \frac{q(a)}{P(a)} - \int_{a,b} q(a)P(b|a) \log P(c = \hat{c}|b) , \quad (74)$$

where we use \cong to indicate equality up to an additive constant. Setting the partial derivatives w.r.t. $q(a)$ to 0 gives

$$0 = \log \frac{q(a)}{P(a)} + 1 - \int_b P(b|a) \log P(c = \hat{c}|b) + \lambda \quad (75)$$

$$= \log \frac{q(a)}{\mathcal{Z}(\lambda)P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c}|b)\right\}} , \quad (76)$$

where \mathcal{Z} is a function of the lagrange multiplier. The result in (70) now directly follows and more specifically the minimizer is

$$q^*(a) = \frac{P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c}|b)\right\}}{\int_a P(a) \exp\left\{\int_b P(b|a) \log P(c = \hat{c}|b)\right\}} . \quad (77)$$

The result in (71) can now easily be obtained by substituting q^* into the KL divergence. \square